Задачи на проверку гипотез

Введение в DS на УБ и МиРА

---

1 Задачи с вариантами ответа (кандидаты в квиз)

1. При проверке гипотезы о равенстве средних p-value оказалось равно 0.04. Выберите верный вариант.
   (a) Основная гипотеза не отвергается на любом разумном уровне значимости.
   (b) Основная гипотеза не отвергается на уровне значимости 1%.
   (c) Основная гипотеза не отвергается на уровне значимости 5%.
   (d) Основная гипотеза не отвергается на уровне значимости 10%.

2. Проверяется гипотеза о равенстве средних против двусторонней альтернативы. Пусть при верной H_0 $Z \sim \mathcal{N}(0,1)$. Чему примерно равно соответствующее p-value, если
   (a) $Z_{obs} = -15$?
   (b) $Z_{obs} = 0$?
   (c) $Z_{obs} = 1.96$?
   (d) $Z_{obs} = 7$?

3. Пусть $X_1, \ldots, X_N$ – выборка независимых одинаково распределённых нормальных величин, $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Предположим, что $N = 1000$. Проверяется гипотеза $\mu = 2$ против двусторонней альтернативы. На выбор есть всего два теста: Z-тест и t-тест. Выберите все верные варианты.
   (a) Тестовая статистика может иметь распределение t_{1000}.
   (b) Тестовая статистика может иметь распределение t_{500}.
   (c) Тестовая статистика может иметь распределение, очень похожее на $\mathcal{N}(0,1)$.
   (d) Тестовая статистика может иметь распределение, очень похожее на $\mathcal{N}(1000,2)$.

4. Определите тип ошибки (I или II рода) по описанию ситуации.
   (a) Мобильный робот врезался в стену. Нулевая гипотеза: впереди нет препятствия.
   (b) Сканер отпечатка пальца не дал согласие на разблокировку системы для зарегистрированного пользователя. Нулевая гипотеза: пользователь есть в базе.

5. Будет ли отвергнута гипотеза о независимости при использовании χ^2-критерия согласия Пирсона на уровне значимости 5%, если

(a) $\chi^2_{obs} = 1$?

(b) $\chi^2_{obs} = 50$?

(c) $\chi^2_{obs} = 100$?

В каждом случае нарисуйте картинку.

2 Задачи с открытым ответом (для тренировки техники проверки гипотез)

1. Рассмотрим выборку независимых одинаково распределённых нормальных случайных величин

$$X = [3, 12, 4, 18, 9, 2, 15],$$
$$X_i \sim \mathcal{N}(\mu, \sigma^2).$$

(a) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 3, \\ H_1 : \mu \neq 3 \end{cases}$$

на уровне значимости 5%.

(b) Постройте 90%-ый доверительный интервал для μ.

2. Рассмотрим выборку независимых одинаково распределённых нормальных случайных величин

$$X = [3, 12, 4, 18, 9, 2, 15],$$
$$X_i \sim \mathcal{N}(\mu, 9).$$

(a) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 3, \\ H_1 : \mu \neq 3 \end{cases}$$

на уровне значимости 10%.

(b) Постройте 95%-ый доверительный интервал для μ.

3. Рассмотрим выборку независимых одинаково распределённых случайных величин $X_1, \ldots, X_{200}$, где $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Оказалось, что

$$\sum_{i=1}^{200} X_i = 20, \ \sum_{i=1}^{200} X_i^2 = 500.$$

(a) Найдите $\bar{X}$ и $\hat{\sigma}^2$.

(b) Проверьте гипотезу

$$\begin{cases} H_0 : \mu = 2, \\ H_1 : \mu \neq 2 \end{cases}$$

на уровне значимости 1%.

(c) Постройте 95%-ый доверительный интервал для μ и при помощи него проверьте гипотезу из предыдущего пункта (уже на новом уровне значимости).

(d) **(пункт со звёздочкой)** Запишите интеграл, при помощи которого можно рассчитать p-value для полученной статистики. Посчитайте этот интеграл при помощи Wolfram Alpha. Убедитесь, что результаты проверки совпадают с предыдущими двумя пунктами.

4. **(хитрая задача)** Рассмотрим выборку независимых одинаково распределённых случайных величин X_1, …, X_{10}. Известно, что распределение X_i не является нормальным. Проверьте гипотезу $H_0 : \mathbb{E}(X_i) = 3$ против двусторонней альтернативы на уровне значимости 5%.

5. **(хитрая задача)** Рассмотрим выборку независимых одинаково распределённых случайных величин Бернулли

$$X = [0, 1, 1, 0, 1, 0, 1],$$
$$X_i \sim \mathrm{Bern}(p).$$

Проверьте гипотезу $H_0 : p = 0.3$ против двусторонней альтернативы на уровне значимости 5%.

6. Рассмотрим выборку независимых одинаково распределённых бернуллевских случайных величин X_1, …, X_{500}, где $X_i \sim \mathrm{Bern}(p)$. Оказалось, что в этой выборке ровно 300 единиц и 200 нулей.

(a) Найдите $\hat{p}$ и оценку дисперсии X_i.

(b) Проверьте гипотезу

$$\begin{cases} H_0 : p = 0.5, \\ H_1 : \mu \neq 0.5 \end{cases}$$

на уровне значимости 5%.

(c) Постройте 99%-ый доверительный интервал для p.

7. Рассмотрим две выборки случайных величин X_1, …, X_{20}, где $X_i \sim \mathcal{N}(\mu_X, 1)$ и Y_1, …, Y_{20}, где $Y_i \sim \mathcal{N}(\mu_Y, 2)$. Будем предполагать, что случайные величины внутри выборок независимы и одинаково распределены, а выборки независимы между собой. Оказалось, что

$$\sum_{i=1}^{20} X_i = 10, \ \sum_{i=1}^{20} X_i^2 = 350,$$
$$\sum_{i=1}^{20} Y_i = 15, \ \sum_{i=1}^{20} Y_i^2 = 400,$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 5%.

8. Рассмотрим две выборки случайных величин $X_1, \ldots, X_{200}$, где $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ и $Y_1, \ldots, Y_{200}$, где $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Будем предполагать, что случайные величины внутри выборок независимы и одинаково

распределены, а выборки независимы между собой. Оказалось, что

$$\sum_{i=1}^{200} X_i = 10, \ \sum_{i=1}^{200} X_i^2 = 350,$$
$$\sum_{i=1}^{200} Y_i = 15, \ \sum_{i=1}^{200} Y_i^2 = 400,$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 10%.

9. Рассмотрим две выборки случайных величин $X_1, \ldots, X_{25}$, где $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ и $Y_1, \ldots, Y_{25}$, где $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Будем предполагать, что случайные величины внутри выборок независимы и одинаково распределены, а выборки независимы между собой. Также будем предполагать, что $\sigma_X = \sigma_Y = \sigma$. Оказалось, что

$$\sum_{i=1}^{25} X_i = 10, \ \sum_{i=1}^{25} X_i^2 = 350,$$
$$\sum_{i=1}^{25} Y_i = 15, \ \sum_{i=1}^{25} Y_i^2 = 400,$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 1%.

10. Рассмотрим выборку объектов из нормального распределения до и после проведения некоторого эксперимента. Выборку ДО обозначим как X, а выборку ПОСЛЕ обозначим как Y. Известно, что

$$\begin{aligned} X &= [10, 15, 20, 18, 15, 20], \\ Y &= [13, 13, 21, 22, 14, 25], \\ X_i &\sim \mathcal{N}(\mu_X, \sigma), \\ Y_i &\sim \mathcal{N}(\mu_Y, \sigma), \end{aligned}$$

Проверьте гипотезу

$$\begin{cases} H_0 : \mu_X = \mu_Y, \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

на уровне значимости 10%.

11. На уровне значимости 10% проверьте гипотезу о том, существует ли зависимость между продолжением образования после окончания школы и типом местности, где выпускник окончил школу. В исследовании принимали участие по 100 школ из каждого типа местности.

	Не продолжил образование	Среднеспециальное образование	Высшее образование
Местность 1	40	40	20
Местность 2	30	50	20

12. **(сложная)** Компания «ГолденАльп» тестирует два новых вкуса шоколада: с орешками и солёной карамелью. Фокус-группа разбивают на две непересекающиеся части: N_1 человек пробуют шоколад с орешками, а N_2 — с солёной карамелью. Каждый участник пробует лишь один тип шоколада и одобряет или не одобряет опробованный вкус. Пусть X_1 — число человек, одобривших шоколад с орешками, а X_2 — одобривших шоколад с солёной карамелью. Будем предполагать, что $X_1 \sim \mathrm{Bin}(N_1, p_1)$, $X_2 \sim \mathrm{Bin}(N_2, p_2)$. Руководство компании «Голден Альп» хочет узнать, есть ли основание полагать, что один вкус шоколада предпочитается другому.

    По результатам эксперимента оказалось, что $N_1 = N_2 = 500$, $X_1 = 400$, $X_2 = 390$. Сформулируйте гипотезу, которая позволит ответить на вопрос компании, и проверьте её на уровне значимости 5%.